SPEECH UNDERSTANDING SYSTEMS

James W. Forgie

Massachusetts Institute of Technology

Prepared for:

Electronic Systems Division
Advanced Research Projects Agency

31 May 1974

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ESD-TR-74-218	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER: AD 783284	
4. TITLE (and Subtitle): Speech Understanding Systems	5. TYPE OF REPORT & PERIOD COVERED: Semiannual Technical Summary, 1 December 1973 – 31 May 1974
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Forgie, James W.	8. CONTRACT OR GRANT NUMBER(s): F19628-73-C-0002
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Lincoln Laboratory, M.I.T., P.O. Box 73, Lexington, MA 02173	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: ARPA Order 2006
11. CONTROLLING OFFICE NAME AND ADDRESS: Advanced Research Projects Agency, 1400 Wilson Boulevard, Arlington, VA 22209	12. REPORT DATE: 31 May 1974
	13. NUMBER OF PAGES: 24
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Electronic Systems Division, L.G. Hanscom Field, Bedford, MA 01730	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

speech understanding systems
linear predictive coding
phonetic recognition
TX-2 system
VASSAL
CASPERS

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The elements of the mid-term system were integrated successfully and it was demonstrated in April 1974. System performance in a controlled test of 110 utterances was 74 percent correct for a 248-word vocabulary, and 50 percent correct for a 500-word vocabulary. Software to support packet speech communication via CVSD modulation techniques over the ARPA computer network has been completed and checked out as far as possible.

This is the last SATS concerning speech understanding systems. Future SATS will be concerned with work on packet speech communication.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

SPEECH UNDERSTANDING SYSTEMS

SEMIANNUAL TECHNICAL SUMMARY REPORT
TO THE
ADVANCED RESEARCH PROJECTS AGENCY

1 DECEMBER 1973 – 31 MAY 1974

ISSUED 23 JULY 1974

LEXINGTON MASSACHUSETTS

SUMMARY

The elements of the mid-term system were integrated successfully and it was demonstrated in April 1974. The system has operated with vocabularies of 248 and 500 words and appropriate context-free grammars to restrict output sentences to those appropriate for the vocal command of our speech data-retrieval, analysis, and display system.

System performance in a controlled test of 116 utterances was 74 percent correct for a 248-word vocabulary, and 50 percent correct for a 500-word vocabulary. Processing time with the smaller vocabulary is on the order of 1 min. for a typical 3- to 4-sec sentence. An experiment is planned to evaluate system performance in greater detail.

The acoustic-phonetic front end has been tested extensively, and substantial modification to the diphthong identification and fricative classification algorithms has been carried out. The performance of our linguistic processing modules has been improved greatly by changes in scoring strategies and by the incorporation of phonologically based and front-end dependent matching rules.

Software to support packet speech communication via CVSD modulation techniques over the ARPA computer network has been completed and checked out as far as possible. This work and the on-line operation of the mid-term system were made possible by the completion of a high-speed data link between TX-2 and the Laboratory's Fast Digital Processor.

CONTENTS

GLOSSARY

APEL	Acoustic phonetic element – output from the acoustic phonetic front end
APEX	TX-2 time-sharing system
ARPA	Advanced Research Projects Agency
CASPERS	A system which processes an APEL string to find and score sentences
CVSD	Continuously Variable Slope Delta modulation – a method of coding speech for digital transmission
FDP	Fast Digital Processor – Lincoln Laboratory computer designed for waveform processing applications
IMP	Interface Message Processor – a computer performing host interfacing and message forwarding in the ARPA computer network
SATS	Semiannual Technical Summary Report
VASSAL	A system which processes an APEL string to find and score sentences

SPEECH UNDERSTANDING SYSTEMS

This is the last SATS concerning speech understanding systems. Work in that area is being terminated at the end of FY 74. Lincoln Laboratory Technical Reports on the Lincoln Mid-term Speech Understanding System and the Speech Data Base are now being prepared which will cover in considerably more detail the work reported here and in previous SATS. Future SATS will be concerned with work on packet speech communication.

I. THE LINCOLN MID-TERM SYSTEM

The elements of the mid-term system were integrated successfully and the system was demonstrated as scheduled in April 1974. A Technical Report is being prepared which will describe the system and its components in detail and will present the results of some systematic testing of its performance. This section presents a brief overview of the system, gives some current performance results, and describes our plans for controlled system tests which are currently getting under way.

A. System Overview

Speech input takes place at a TX-2 console using a close-talking, noise-canceling microphone. The speech is digitized in TX-2 and sent via a 1.6-megabit serial connection to the Fast Digital Processor (FDP) where parameter extraction and phoneme class segmentation take place. The results are returned to TX-2 where the remainder of the phonetic recognition is carried out to yield a string of acoustic-phonetic elements (APELs). These serve as input to a linguistic processing module which uses a context-free grammar and vocabulary appropriate to a task domain to find and score sentence candidates. Output of the system is a display of the sentence candidates (if any) which the linguistic module succeeded in matching to the APEL string. When the decision was made to terminate the project, work on the functional response module discussed in the previous SATS[1] was stopped prior to implementation. The system, therefore, lacks a functional response output and is more properly termed a speech recognition system than a speech understanding system.

The system makes use of the phonetic recognition processing discussed in Sec. II of this report and in previous SATS[2] and publications.[3,4] It is compatible with both the CASPERS and VASSAL linguistic processing modules discussed in Sec. III of this report, in the previous SATS (see pp. 11-15 in Ref. 1), and in a recent publication.[5] To date, all on-line experiments and demonstrations have made use of the VASSAL linguistic module, and the results discussed in this section apply to that version of the system.

The system is capable of recognizing sentences from any subset of English for which a context-free grammar and vocabulary are available in the proper form. Currently, we have three different grammar and vocabulary combinations which can be demonstrated. Two are appropriate to the proposed Lincoln task of vocal command of our speech data retrieval, analysis, and display system. The other is similar to the System Development Corporation (SDC) task of querying a submarine data base in a formal query language. The two Lincoln tasks differ in vocabulary size and syntactic complexity. The smaller of these, which has been used in most of our demonstrations and tests so far, has a vocabulary of 248 words and a finite

gramma r capable of generating 4.8 million sentences. The larger task has a vocabulary of about 500 wor ls and a grammar that will allow something in excess of 500 billion sentences.

mall Lincoln grammar allows only command sentences which can begin with any one of nand verbs. The grammar allows 15 sentence types which constrain the objects to one: atically appropriate to the command verbs. Modifiers are similarly constrained. The l. .g grammar allows some questions as well as commands. It handles 120 command verbs and five q estion words.

The following are sample sentences acceptable to the small grammar:

"List the back vowels from utterance one hundred seven."
"Recompute the average energy in the second voiced segment."
"Set the default for sex to male."
"Move to the next utterance."
"Drop the edited labels from the drum."
"Display he formant graph on the Hughes scope."
"Erase the display of the confusion matrix."
"Put the right cursor on the third frame."
"Connect tape unit four to console number one."
"Calculate the effec. of setting the threshold to one."
"Delete those greater than six seconds."
"Go into the graphics mode."
"Search for the twenty kilohertz waveform."
"Skip to the third sentence on tape unit six."
"Give me the range of the third formant."
"Calculate the distribution of the liquids."

In addition to the above, the large grammar will accept sentences such as the following:

"Now please put up the distribution of energy for the example."
"Edit the phonemic transcription of this statement."
"I want to see the ninth sample by a female speaker."
"Go on to the final entry with a liquid-fricative occurrence."
"Shift the label ahead 15 milliseconds."
"Select the average spectra for the fricative example in each statement by speaker R.W."
"Put the phonemic labels under the spectrogram."
"How many statements have parse-trees in the data base?"
"What is the owner's name?"

Processing time varies with sentence duration, grammar size and complexity, and difficulties encountered in recognizing individual utterances. When using the small (248-word) Lincoln grammar and the VASSAL linguistic processing, a typical 3- to 4-sec sentence will require on the order of 1 min. to complete all processing if there is no other time-sharing activity on TX-2. This processing time is short enough to allow a substantial number of sentences to be processed by the system for evaluation purposes.

B. System Performance

During the past several months, we have processed about 375 sentences which were acceptable to the small (248-word) Lincoln task grammar. About 40 speakers have used the system, many speaking only one or two sentences during a demonstration. Two speakers have spoken about 100 sentences each. These sentences also are acceptable to the large (500-word) grammar, but only about one-third have been processed as yet using the large grammar. We have processed only a few sentences with the SDC submarine data base query language, and we will not report results with that grammar at this time.

Our phonetic recognition processing contains some speaker-dependent decision mechanisms. The vowel formant space is normalized to the individual speaker from calibration sentence data. The fricative recognition algorithm uses thresholds which are different for male and female speakers. Adjustment of these thresholds is a manual procedure which is not practical during demonstration situations. Therefore, in demonstrations we have used one or another of the preset speaker profiles which we have generated for the Lincoln speakers who account for most of the data. This procedure has worked quite well for male speakers, often achieving success on voices which sounded quite different from the calibration speaker. Successful recognition was achieved in several cases where the speakers had pronounced foreign accents.

Since the Lincoln Speech Data Base is an integral part of our system, all system demonstrations automatically produce data for the data base. Data collected in this way provide a diverse but uncontrolled sample, and there are many artifacts such as incorrect microphone positions, and extraneous background noises which make the total sample a poor one for measuring performance. We are preparing a new data sample for a controlled system experiment, and Sec. C below describes our plans in that area. The data we have now, on which we feel it is reasonable to make comparative judgments, are a set of 116 sentences spoken by 6 male speakers for whom we have speaker normalization profiles and which are free of artifacts.

For these 116 sentences, using VASSAL linguistic processing and the small grammar, the system correctly recognized 74 percent of the sentences. About half of the incorrect sentences were wrong in only one word. In one case, no sentence at all was found. In ten cases, a completely wrong sentence was found.

For the same sentences and VASSAL processing, but using the large grammar, the system correctly recognized about 50 percent of the sentences. Again, about half of the incorrect sentences were wrong in only one word.

It should be emphasized that these results are obtained from a small sample of the possible sentences which the system could be expected to handle. Because the APELs generated by the phonetic recognition module have some inherent ambiguity as well as potential for error that is not randomly distributed, the performance on any test will depend on the particular sentences chosen for testing. While these results correspond to a particular test set, we feel that they are representative of the results the system has achieved in live demonstration situations. It would be possible, however, to select sentences which either probe the weakness of the acoustic processing or explore the regions of the sentence space for which the grammar provides less useful constraint. The techniques we have been using to describe the task domain, so that speakers can produce acceptable sentences, tend to produce a sentence set that is well distributed with respect to the semantics of the task domain. This distribution is not uniform with respect to sentence difficulty, and we expect that test sentences so selected would result in lower overall performance.

C. Evaluation Experiment Plans

While experience during the past several months has given us a reasonable idea of the success and limitations of the current system, we feel that a well-designed experiment is needed to (1) obtain some measure of acoustic-phonetic processing performance, (2) obtain some measure of linguistic processing performance, and (3) attempt to identify areas where change will lead to overall system improvement. Our proposed design for this experiment follows.

1. Subjects

We can predict to some degree how successful our system will be with a particular speaker. First, we do much better with male than with female speakers. In part, this is due to a few known and easily adjusted thresholds. Whether the adjustment of these thresholds will lead to a level of performance that will permit meaningful analysis is not known at this moment. We do plan to investigate this and include whatever seems reasonable in our report.

We know that some speakers tend to place a great deal of emphasis on the first word or two in a sentence and then let the rest of the sentence trail off dramatically. We will not include any speaker who exhibits this tendency in pretesting. We further know that speakers with a rather rich tonal quality and even cadence tend to be successful. This does not imply that we will run many subjects and only include for analysis those with good results. Rather, the idea is to use some known speakers and choose others from knowledge of their speech, and listen to them read a few sentences.

We will use six male speakers - three known and three new. We will include female speakers contingent upon the success of the adjustments mentioned above. At a minimum, we will attempt to analyze our problem with female speakers.

2. Sentence Material

Most of the sentences will be acceptable to the small Lincoln grammar. Some will be acceptable only to the large grammar. Each sentence will be formulated from a chart similar to the one in Fig. 1 which represents a small part of the grammar, and is rich enough to provide a variety of sentences, but is restricted enough to be understandable. The entire small grammar is spanned with 29 such charts.

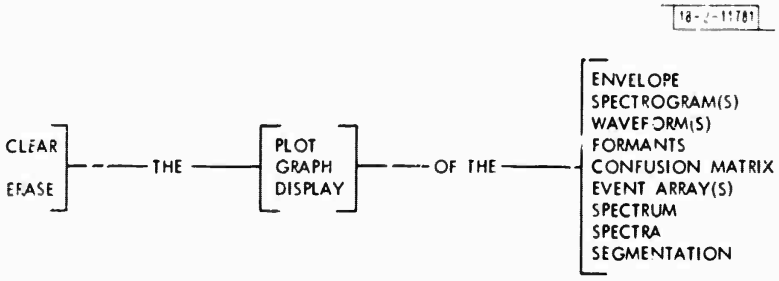

Fig. 1. Typical chart to be used in generating sentences for system evaluation.

3. Method

Each speaker will begin by saying two calibration sentences. Data from these sentences will be entered for the speaker. A description of the task will be provided and a presentation of the vocabulary given to avoid difficulties during sentence formulation. The speaker will be shown one of the charts and guided through the formulation of several sentences. Once there is assurance that this aspect of the task is understood, the speaker will say a few trial sentences for interpretation by the system. Following the resolution of any apparent difficulties, the speaker will be presented with a random ordering of other charts and asked to formulate one sentence from each chart. The speaker will be asked to try to make some short, some moderate, and some long sentences. The experimenter will caution the subject if the sentence is not within the grammar. In the event of an illegal sentence, the nature of the problem will be pointed out and a new sentence will be formulated.

About a week after the initial sentences have been spoken, the speaker will return and reread the calibration sentences, and read a scrambled ordering of the legal sentences he formulated earlier.

4. Analysis

The calibration sentences will be used to obtain vowel formant frequencies and fricative energy for each speaker. These normalizing values will be entered before any sentence processing.

The only parameters for the acoustic-phonetic processing are a male/female threshold for fricative identification and the normalizing data from the calibration sentences. The linguistic processing makes use of speaker independent data derived from experience with sentences processed prior to the experiment. These latter data consist of a matrix of phoneme/APEL crosstabulations, and rules used to edit the APEL string. The crosstabulation matrix is based upon about 100 sentences from about 10 different speakers. The rules for APEL editing have evolved over a period of 6 months from experience with some 200 sentences.

Data from each speaker will be considered independently. The analysis involves several distinct and independent phases.

First, we will look at the total system performance. The primary results will be the proportion of correct sentences and an analysis of errors. Generally, we find sentences to be either correctly recognized, wrong in one or two words, or completely wrong. Where the sentence is wrong in one or two words, the error typically occurs where there are many choices for the expected construct. For example, if a number is expected, "sixteen" may be recognized when, in fact, "sixty" was spoken. Completely wrong sentences generally result from missing the first word, thus leading to utter deterioration.

In this first phase, we will consider the differences between the formulated and the read sentences. While there may be some dispute about the free speech component of sentences made from our charts, it is our contention that the process is quite different from reading.

The second phase of analysis involves the generation of a phoneme/APEL crosstabulation. To achieve this, we will give the linguistic processor the APEL string and the correct sentence. The processor makes an optimal match, considering phonological rules, and fills in entries in the matrix. The match is constrained to never produce a vowel/consonant entry and vice versa.

While this clearly differs from the process previously employed in evaluating acoustic-phonetic processing, a rough comparison indicates that the results are similar. The purpose of this phase of analysis is to obtain further data on our acoustic-phonetic processing.

The third phase is directed toward attempting to identify some areas in which further research will lead to progress in speech recognition. While the problem of speaker normalization is accepted as a challenge, our experience indicates that the entire system is reasonably robust and not heavily dependent upon obtaining normalizing values for each speaker. One possible reason for this is that we simply have not addressed the right aspects of the problem. At a minimum, we plan to compare the normalizing values for the repeated calibration sentences with the original. We also plan to compare these values across speakers and hope to get some idea of how sensitive the system is to variations in these parameters. We further plan to examine differences in the crosstabulation matrices among speakers and counterposed to the matrix used in the recognition phase.

The scope of our analysis is heavily dependent upon the use of our data base. Where appropriate, we are able to look at waveforms, formants, spectra, etc. in an attempt to understand the behavior of the acoustic-phonetic processing. Even after the experiment has been reported, we will have a rational corpus of data to explore our ideas concerning improving performance.

II. PHONETIC RECOGNITION

Extensive tuning and testing of the acoustic-phonetic front end, as well as a few substantial modifications, have been carried out since the last SATS. Fairly complete documentation of the front end and its performance is now available in published form,[1] and additional documentation will be included in a forthcoming Technical Report. The remainder of this section describes the algorithms for diphthong identification and fricative classification, both of which have undergone substantial changes since the last SATS. The experimental results given are from a corpus of 111 sentences, distributed among 7 speakers (6 male, 1 female).

A. Diphthong Identification

Every VWL segment (representing a detected vowel nucleus) is subjected to special tests to detect the presence of the diphthongs /ɑy/* (buy) and /ɔy/ (boy), and the diphthong-like sound /yu/ (compute). Detection of a particular diphthong occurs when the formant trajectories in the VWL meet a set of criteria peculiar to that diphthong. The overall performance of these algorithms is quite good, with over 80-percent correct identification and false alarms in less than 1 percent of the remaining syllables.

The /ɑy/ detection algorithm is illustrated in Fig. 2. We require that the total duration T_v of the VWL exceed a threshold, that there be a region greater than $T_v/3$ in length where the slope of F2 is positive, and that the formant positions at a point $T_v/4$ from the left edge of the VWL be characteristic of a steady state /ɑ/. The test for this third requirement is carried out by comparing measured F1 and F2 with tabulated values for /ɑ/ in a speaker-normalized vowel table, in a manner similar to that described previously for steady-state vowel identifications. Then we require either that the rise in F2 exceeds 300 Hz or that the duration of the

* In this report, phonemes are represented in standard International Phonetics Association notation and are marked with slashes (e.g., /ɑy/); APELs are defined as introduced and are marked with braces (e.g., {F}).

owel exceeds 250 msec. This last duration test was helpful in detecting /ɑy/'s in words like "file," "time," or "five" where the influence of the following consonant caused the rise in F2 to be less than in other situations, but where it was found that the durations T_v and T_{pos} were longer than average. As seen in Fig. 3, this algorithm correctly detected 80 percent of the /ɑy/'s, with half the misses detected as /ɔy/. There were a few false detections, mostly with vowel and glide combinations like /le/ which have somewhat similar formant trajectories.

The /ɔy/ detector employs a strategy very similar to the /ɑy/ detector. F1 and F2 near the beginning of the VWL are required to reach positions close to those of a steady state /ɔ/, and a sufficiently long region of increasing F2 is required, with a change in F2 of at least 750 Hz. There were only 9 /ɔy/'s in the 411 sentences, and all were correctly detected. In three cases, the diphthong /ɑy/, preceded by /w/ or /r/, was detected as /ɔy/; and there were three additional false alarms.

The /yu/ detection algorithm applied a technique similar to the one previously described for /r/ detection.[2] A set of measurements are thresholded and any that exceeds a threshold contributes a penalty to a score proportional to how much it exceeds the threshold. The requirements are that F1 be low throughout the segment, that F2 be high on the left and have negative slope over at least 1/3 of the segment, that F3-F2 become small near the middle of the segment,

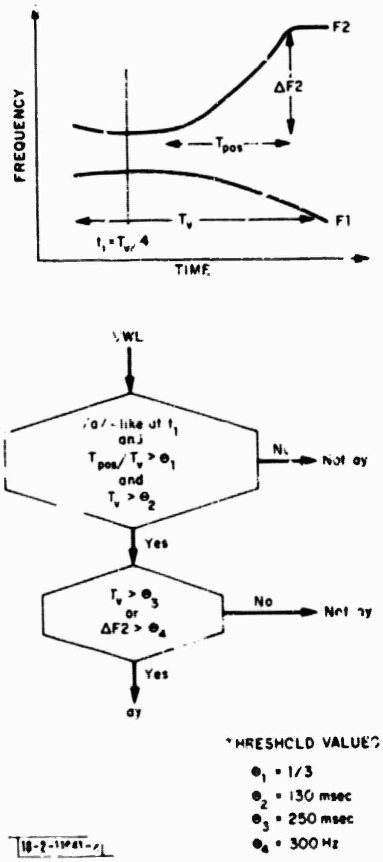

Fig. 2. Algorithm for identification of diphthong /ɑy/.

18-2-11642-2

HAND-LABELED PHONEMES \ APELs	ɔy	ɑy	yu	OTHER
/ɔy/	9			
/ɑy/	3	24		3
/yu/			16	5
OTHER VOWELS	3	5	3	931

Fig. 3. Confusion statistics on diphthong identification.

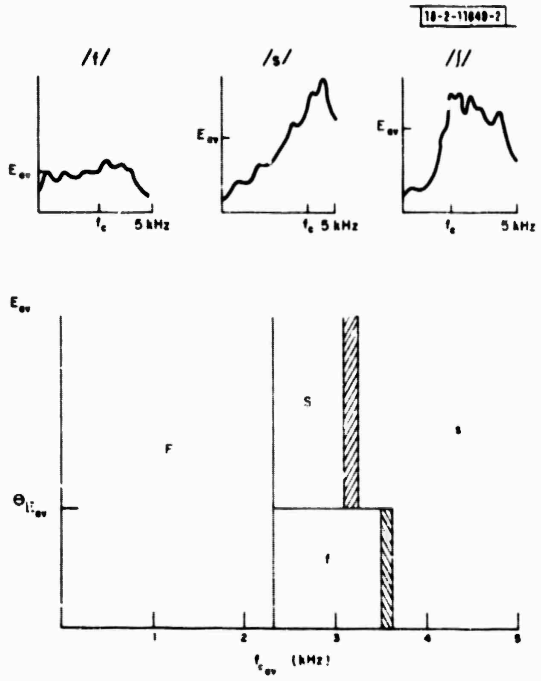

Fig. 4. Algorithm for fricative classification.

Fig. 5. Confusion statistics on fricative classification.

18-2-11648-2

HAND-LABELED PHONEMES \ APELs	s	f	S	F
/s/,/z/	202	7	4	1
/f/,/v/,/θ/	8	64	3	1
/ʃ/,/ʒ/,/tʃ/,/dʒ/	4	2	26	
/h/	1	4	1	1
ASPIRATION AFTER VOICELESS STOPS /p/,/t/,/k/	9	18	7	
VOWELS		2	3	4

and that F3 decrease sharply from the beginning to the middle of the segment. The results are indicated in Fig. 3; 16 of 21 /yu/'s were detected, with five false alarms.

B. Fricative Classification

The basic fricative classification algorithm takes all FRIC segments[2] and categorizes them into phoneme classes as follows: the APEL {f} designates the phonemes /f/ or /θ/, the APEL {s} designates /s/ or /z/, the APEL {S} designates /ʃ/ or /ʒ/, and the APEL {F} indicates that the FRIC segment did not meet the criteria for any of the above phonemes. The primary parameters used are segment duration, a measurement E_{av} of the volume [RMS (0-5000)] averaged over 5 frames at the segment center, and a critical frequency $f_{c_{av}}$. (These measurements are different from those used in the fricative classification algorithm reported in the last SATS.) With S(f) representing the spectral amplitude in a particular frame, and $\Delta f = 10000/256$, we define f_c for that frame as $f_c = k_c \Delta f$ where k_c is the largest integer for which

$$\sum_{k=k_c}^{k=127} S(k\Delta f) \Big/ \sum_{k=0}^{k=k_c-1} S(k\Delta f) \geq \theta_c \quad .$$

In these experiments, θ_c was chosen as 1/2, so that f_c corresponded to the center of mass of the spectrum. To determine $f_{c_{av}}$, we average f_c over 5 frames at the center of the segment. The fricative classification algorithm has the option of segmenting the FRIC segment into two APEL segments, if distinct contiguous fricatives are detected (e.g., "voiceless fricative").

If the FRIC segment is less than 100 msec in duration, no additional segmentation is attempted. The FRIC is classified according to E_{av} and $f_{c_{av}}$ as indicated in the bottom half of Fig. 4. (In the overlap regions, two APEL choices are given.) The frequency regions, defined on the figure, are used for male speakers; a different set of regions is needed for female speakers. The volume threshold T_a was empirically determined. The basic properties being exploited are illustrated by the spectra sketched in the top half of Fig. 4. The fricatives /f/ and /θ/ are weak and have relatively flat spectra; the spectral energy for /s/ and /z/ is skewed toward higher frequencies than for /ʃ/ and /ʒ/. A spectrum skewed too much toward lower frequencies leads to a suspicion that a segmentation error may have been made and that the segment may be vowel-like. If the FRIC segment is greater than 100 msec in duration, then separate classifications are made of the left and right halves of the segment. If these classifications are different, the FRIC segment is split into two APELs. Results of fricative classification are summarized in Fig. 5. The statistics indicate 91-percent correct classification, as determined by summing the diagonal elements in the first three rows and dividing by the sum of all elements in the first three rows. A few /v/'s which were called {f}, and /tʃ/'s and /dʒ/'s called {S}, were included as correct classifications.

A number of stop aspirations were segmented as FRIC; they were further classified by the fricative algorithm as indicated. Of the few FRICs produced due to voicing errors in vowels, about half were assigned APEL {F} because of low $f_{c_{av}}$. The double-fricative detector performed successful segmentation and identification of 83 percent of adjacent but distinct fricatives present in the spoken sentences. (These results represent about 20 detections of double fricatives; in each case where correct segmentation was performed, the identification was

also correct.) We are working on the problem of distinguishing /z/ from /s/ and /ʒ/ from /ʃ/ on the basis of the pitch detector voicing indicator. Some promising results have been obtained, but difficulties are caused by the fact that many /z/'s are devoiced in continuous speech, at least over part of their duration.

III. LINGUISTICS

A. VASSAL

The accumulation of a substantial number of processed sentences has led to a clearer understanding of the basic patterns of the acoustic-phonetic front-end errors. The changes to the matching and scoring algorithms which were indicated by these data were implemented, and these changes raised considerably the level of performance of the system.

The most direct use of these new data was in the automatic generation of the scoring matrices. For each sentence in the data base, there is an associated lexical transcription and the sequence of APELs produced by the front end. An "automatic labeling" program was used to align the phonemes in the nominal pronunciation with the APELs. From this correspondence, two confusion matrices were built up(one for vowels, one for consonants). The scoring matrices were then computed by taking the LOG_2 of the estimated probability that a particular APEL is produced when a given phoneme is spoken. An increment was added to make 0 equivalent to the LOG_2 (1/5). A few (less than 2 percent) of the scores were manually edited to compensate for inadequate data in the sample.

Also, a basic change in the matching mechanism involved an improved solution to the so-called "alignment" problem. This problem occurs because of the front-end (or speaker) errors that produce missing and spurious segments in various contexts. The revised scheme takes advantage of the fact that vowels usually contain more energy than consonants and are much less likely to be missed. The relatively few missing or spurious vowel segments often can be predicted, thus reducing even further the possibility of error. When matching, VASSAL aligns the words with the APELs by first lining up the vowels. This considerably simplifies the matching since we have reduced the problem to one of aligning small sets of consonants with small sets of APELs.

A third significant change was the refinement of the evaluation function which is used to score each partial sentence candidate produced thus far. This ensures that at any point in time the most promising hypothesis will be pursued. Since not all fragments are of the same length, some technique must be used to normalize them. One simple scheme is to use the average score per APEL "consumed." However, a slightly more complex scheme allows some flexibility in varying the behavior of the system. In this scheme, we have two components to the evaluation function. To the total score obtained thus far by matching, we add a predicted score for each unconsumed APEL. By being optimistic in the prediction, we favor the shorter candidates and will produce slower, but more accurate, behavior (if we don't run out of space or time); by being pessimistic, we produce faster, but probably less-accurate, results. The effect of this component becomes less important as we process more of the sentence and there are fewer unconsumed APELs.

Much of the information concerning the mapping between the phonemes in the dictionary words and the APELs in the input is embodied in three sets of rules. One set is used to edit

the sequence of APELs. Those segments almost certain to be spurious (such as a voiced silence following a nasal) are deleted, and others are flagged as possibly spurious.

The second set contains those rules which predict possible spurious segments based on phonemic context. For example, the aspiration or burst of a stop consonant is sometimes marked as a fricative. Thus, an optional {f} is inserted in those words containing prestressed stops. This second set also maps members of those sets of phonemes which are not differentiated by the front end (e.g., /s/ is not distinguished from /z/) into a single element.

The rules in the first two sets are applied to the APEL string before matching is attempted, and to the dictionary at compile time. However, the third set contains rules which are functions of both phoneme context and corresponding APELs, and these are applied at run time. Many of these rules pertain to the influence of neighboring semivowels on vowel classification For example, an /r/ preceding or following a vowel may cause it to be classified as an {R} (the vowel of "bird"). Rules pertaining to word boundary behavior are also included in this set.

Some of the rules in the three sets represent well-known phonological phenomena, such as the optional insertion of /t/ between /n/ and /s/ (e.g., "console"). However, many of these rules are artifacts of the particular segmentation and classification algorithms now being used by the front end, and could not be predicted in advance. The collection of a substantial number of processed sentences was necessary to distinguish between recurring patterns and isolated examples caused by such variables as noise or formant tracker error.

B. CASPERS

CASPERS is the name given to a system being developed as part of the ongoing doctoral research of John W. Klovstad at M.I.T. Its general operation was described in the last SATS (see pp. 11-15 in Ref. 1), and its extension to handle phonological rules which apply at word boundaries was described in a paper for the IEEE Symposium on Speech Recognition.[5] Recent work has been concerned with changes in scoring strategy and the accumulation of phonological and front-end dependent rules and their incorporation into the matching process.

CASPERS now uses two distinct scoring strategies. As before, an average score per APEL consumed is used to select the parse path to be pursued next, but in the word-scoring process, an accumulated score is used which favors longer words. In this word-scoring process, a scoring matrix derived from experience with the system supplies scaled log probability values which are then offset with a dynamically adjusted value depending upon the overall average path score.

The process of tuning and testing CASPERS has proceeded by deriving statistics from most of the sentences accumulated in the tests and demonstrations of the Lincoln system which were discussed in Sec. I of this report. From this experience, something in excess of 60 matching rules has been incorporated into CASPERS. Roughly, two-thirds of these have a phonological basis. The remainder are front-end dependent rules introduced to account for the behavior of the particular front-end processing in the Lincoln system. A recent fine-tuning pass over 70 sentences found only one or two which could not be recognized successfully after the fine tuning. The system is now about ready for testing with new speech material.

IV. PACKET SPEECH COMMUNICATION

Lincoln Laboratory is scheduled to participate in an experiment to send and receive speech over the ARPANET using the CVSD (Continuously Variable Slope Delta) modulation technique.

Other participants in the experiment are expected to be the Information Sciences Institute of the University of Southern California (ISI) and the Speech Communication Research Laboratory at Santa Barbara, California. The CVSD technique gives tolerable quality of speech transmission in the 8- to 20-kilobit per-second range. The purpose of the experiment will be to determine to what extent this technique can provide acceptable speech under packet switched network conditions where there is appreciable dispersion of the transmission delays across packets.

The remainder of this section describes the CVSD modulation technique and the software we have generated for TX-2 and the FDP to carry out the experiment. We have tested the system to the point of successfully looping coded speech through the Lincoln IMP and reconstituting the speech at the FDP. Further testing awaits completion of similar software at the other sites.

A. CVSD Algorithm

The basic CVSD algorithm is outlined in Fig. 6. Both the transmitter and the receiver generate an estimate for each speech sample, s'(n), by an identical process. The transmitter compares each new speech sample s(n) coming in from the A/D converter with the previously generated estimate, s'(n – 1), and sends across the channel and to the M analyzer a bit to indicate which is larger [1 if s'(n – 1) > s(n); 0 if s'(n – 1) < s(n)]. Therefore, a bit stream comes across the channel, one bit per sample, and the bit rate is equal in bits to the sampling rate in hertz.

The function of the M analyzer is to compute a quantity M(n) such that $M(n) + a_1 s'(n-1)$ will be a satisfactory estimate for s'(n). To do this, the M analyzer uses the three most-recent

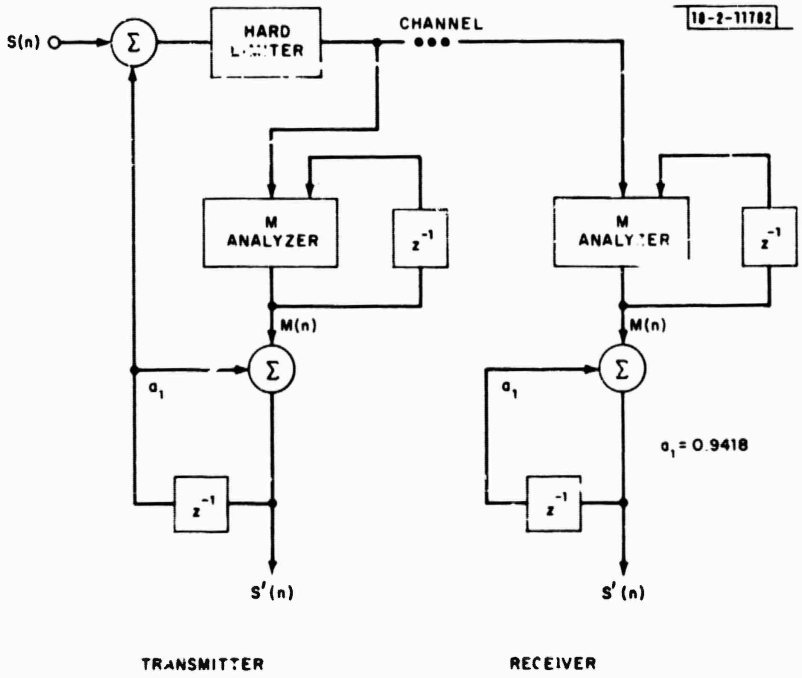

Fig. 6. Flow chart for CVSD modulation algorithm.

bits, generated from s(n), s(n – 1), and s(n – 2), to determine whether the magnitude of M should be increased or decreased (see Fig. 7). The current bit b(n) determines the sign of M(n) (1 = negative). |M| is increased whenever the last three bits are identical (either 111 or 000). If there is a string of 0's, it means that s'(n – 1) was less than s(n) for three samples in a row, and therefore that M should be increased. Likewise, three 1's mean that s'(n) was greater than s(n) three times in a row and, therefore, that a larger number should be subtracted or that |M| should be increased. |M| is increased along a curve which asymptotically approaches a limit MMAX, and is decreased along a different curve asymptotically approaching MMIN.

Silence detection is also dependent upon the pattern of the last three bits, or rather, upon a history of 3-bit patterns. The most likely pattern of bits for silence is alternating 0's and 1's, whereas, in non-silence it is likely to frequently encounter a string of 1's or a string of 0's. A running score is accumulated based on this assumption, as indicated in Fig. 8. If this score remains in the "silence" range for a sufficiently long time (currently 200 msec) then a bit indicating silence is turned on. It remains on only until the first time that the score falls outside the silence range.

There is a delay of 100 msec between what is sent out and what is computed. This delay is beneficial both in going into and coming out of silence. It is necessary to detect at least 200 msec worth of silence before indicating silence in order to be sure that the silence is a pause and not just a stop gap. The 100-msec delay allows silence to be indicated for the second half of that time. In coming out to silence, several milliseconds of speech often pass by before the running score actually succeeds in getting out of the silence region. The delay now allows those samples to be called speech, by turning off the silence indication 100 msec before speech was first detected.

B. TX-2 Packet Speech Software

A TX-2 program to handle packet speech over the ARPA network has been written. The program follows a nonstandard protocol proposed by Mr. Danny Cohen of ISI. This protocol allows the sender and receiver to establish a link between two computers on the network and then to send and receive standard size packets of speech. Control messages are also specified which notify either host that the other's buffers are full and that the sending must be temporarily halted. A method of ending transmissions is also provided.

The protocol was implemented on TX-2 by bypassing the standard Network Control Program; therefore, no other network traffic is honored while the Packet Speech Program is in operation. The program runs as a part of APEX and does not interfere with normal time-sharing operations.

TX-2 acts as a source of buffer space and dispatcher. A word in each direction is exchanged with the FDP over the data link described in Sec. V. The rate of exchange is determined by the FDP CVSD algorithm. Speech from the FDP is buffered up and then sent out over the network. Speech from the network is buffered up and then sent to the FDP when a sufficient amount has accumulated. The amount received before sending to the FDP is an experimental parameter which we intend to explore in order to minimize the "glitches" caused by random network delays.

V. SYSTEM ACTIVITIES: TX-2/FDP DATA LINK

In order to support on-line operation of the mid-term speech understanding system and the packet speech experiments, a high-speed serial data link connecting IO channels on the two

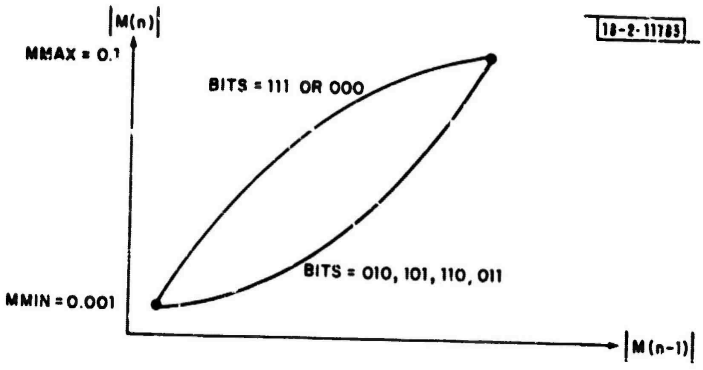

Fig. 7. Operating characteristic for M analyzer in CVSD modulation algorithm.
If bits = 111 or 000:

$$|M(n)| = k[|M(n-1)| - MMAX] + MMAX \quad .$$

Otherwise:

$$|M(n)| = k[|M(n-1)| - MMIN] + MMIN$$

where k = 0.988889.

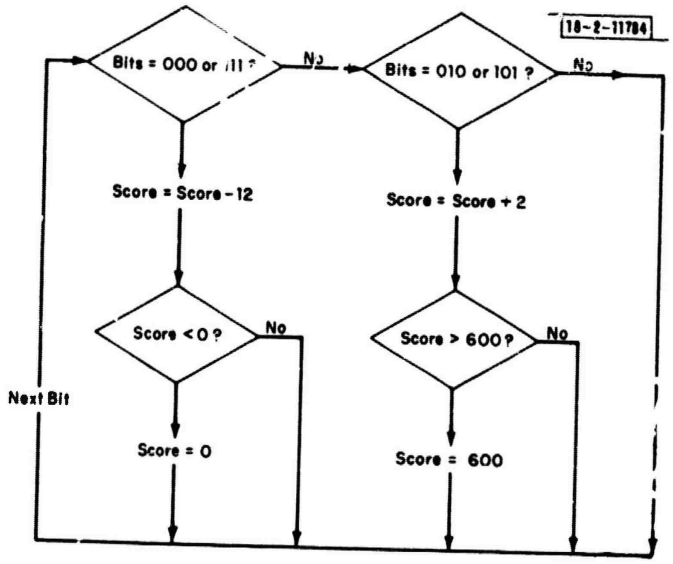

Fig. 8. Flow chart for CVSD silence detection algorithms. Score > 470 = silence.
Bits = last three bits sent across channel.

machines has been installed and is now operational. The hardware consists of digital interface circuits and a pair of frequency shift serial bit stream modems.

The digital interface circuits at each machine provice proper data buffering and handshaking with their respective machines. They also provide parallel-to-serial conversion from machine data word output to serial modem input, and serial-to-parallel conversion from modem receiver output to machine input. Since the TX-2 data word is 36 bits and FDP is 18 bits, the FDP transmits and receives pairs of 18 bits each which correspond to single TX-2 words.

The modems operate as frequency shift modulators and demodulators. A serial data stream at 600 nsec/bit into the modem transmitter causes an output of 13.3- or 16.7-MHz bursts, each lasting 600 nsec. Before a serial word is output to a modem, a frame pulse of 600 nsec precedes it and generates a third frequency burst of 20 MHz. The transmitter sequence is as follows: first a 600-nsec burst of 20 MHz for the frame pulse, followed by eighteen 600-nsec bursts of 13.3 or 16.7 MHz depending on data zeros or ones, and a final control bit of 600-nsec duration also 13.3 or 16.7 MHz. The transmitter drives an RG-22B balanced cable with these frequency bursts. At the modem receiver, the RF signal is filtered with filters matched to the three 600-nsec frequency bursts. The filter outputs are envelope-detected and drive differential comparators. The comparator driven by the 13.3- and 16.7-MHz filter envelopes produces a data bit stream. A pair of comparators produces the frame bit preceding the valid data stream. In addition, a phase lock loop and voltage controlled oscillator produce a serial clock suitable for driving the digital interface. This clock, data stream, and a frame pulse stream are three separate output lines which drive the digital interfaces.

REFERENCES

1 Speech Understanding Systems Semiannual Technical Summary, Lincoln Laboratory, M.I.T. (30 November 1973), pp. 19-20, DDC AD-774452/7.

2. Ibid., pp. 3-11; see also Speech SATS (30 November 1972), pp. 1-6, DDC AD-754940 and Speech Understanding Systems SATS (31 May 1973), pp. 1-6, DDC AD-763723.

3. C. J. Weinstein, S. S. McCandless, L. F. Mondshein and V. W. Zue, "A System for Acoustic-Phonetic Analysis of Continuous Speech," Proceedings of IEEE Symposium on Speech Recognition (IEEE Catalog No. 74CH0878-9 AE), April 1974, pp. 89-100.

4. S. S. McCandless, "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," IEEE Trans. Acoustics, Speech, and Signal Processing ASSP-22, 135-141 (1974).

5 J. W. Klovstad and L. F. Mondshein, "The CASPERS Linguistic Analysis System," Proceedings of IEEE Symposium on Speech Recognition (IEEE Catalog No. 74CH0878-9 AE), April 1974, pp. 234-240.